4-2018

Sum frequency generation spectroscopy of simulated protein secondary structures.

Andrew J. Adams
University of Louisville

SUM FREQUENCY GENERATION SPECTROSCOPY OF SIMULATED PROTEIN SECONDARY STRUCTURES

By

Andrew James Adams
B.S., University of Louisville, 2017

A Thesis
Submitted to Faculty of the
J.B. Speed School of Engineering of the University of Louisville
In Partial Fulfillment of the Requirements
for the Degree of

Master of Engineering
in Chemical Engineering

Department of Chemical Engineering
University of Louisville
Louisville, Kentucky

May 2018

SUM FREQUENCY GENERATION SPECTROSCOPY OF SIMULATED PROTEIN SECONDARY STRUCTURES

By
Andrew James Adams
B.S., University of Louisville, 2017

A Thesis Approved on
April 18, 2018

by the following Thesis Committee:

_____

Thesis Advisor
Dr. Vance Jaeger

_____

Dr. Thomas Starr

_____

Dr. Eric Rouchka

ABSTRACT

SUM FREQUENCY GENERATION SPECTROSCOPY OF SIMULATED SECONDARY STRUCTURES

By
Andrew Adams

April 18, 2018

Sum frequency generation (SFG) spectroscopy is an experimental technique for differentiating between various conformations and orientations of interfacial proteins. Combining a theoretical framework for SFG with molecular dynamics (MD) simulations provides a powerful tool for studying systems containing interfacial proteins with applications in cell transport, biofilms, and fermentation processes. Roeters' method was used to calculate theoretical SFG responses for a variety of individual α-helix and β-sheet peptide secondary structures simulated using MD. Results show how the shape and locations of SFG amide I responses change with differences in hydrogen bonding patterns, peptide orientations, and SFG polarization combinations. The data presented herein demonstrate the utility of SFG spectroscopy for uniquely describing the orientation and conformation of interfacial proteins and how molecular simulation and theoretical spectral calculations complement this experimental technique.

TABLE OF CONTENTS

LIST OF FIGURES

INTRODUCTION

General

Proteins preform an immense number of functions that are essential to life as we know it. A large majority of vital cell functions are carried via the use of proteins whether that be by sending signals throughout the body, fighting infections, transporting nutrients, or catalyzing biochemical reactions. Since proteins are a crucial part of most biological functions, they have long been a topic of interest to researchers across many disciplines of science; however, researching proteins have proved challenging due to their large, intricate, and complex nature. Various experimental methods for characterizing the structures of proteins have been used throughout the years such as x-ray crystallography, NMR spectroscopy, and cryo-electron microscopy. These methods have the drawback of needing to remove the protein from its natural solvated state and being placed into a vastly different environment than the protein's native one. Additionally, while these methods allow for data to be gathered on the components and structures of proteins, they largely lack the ability to do so *in situ*. Since the environment and surroundings of a protein greatly affect the structure of the protein, these methods can fall short in giving useful information in situations where the interfacial structure and properties of proteins are of interest (i.e. biomaterials, biofilms, lipid membranes, etc.). Currently, none of the over 100,000 protein structures stored in the Protein Data Bank were obtained from proteins while at an interface. A form of vibrational spectroscopy called sum frequency generation (SFG) is able to provide information on interfacial protein structures. This utility of SFG stems from

it being a second order non-linear optical spectroscopy which is also why SFG is only able to be used for characterization at interfaces. There is also the problem that the data generated from SFG experiments gives only general data about the secondary structure and orientation of a given protein and does not provide information on the primary structure of the protein. One way to tackle this problem is through the use of computers and molecular dynamics simulations, which provide a physics-based model of hypothetical protein structures. With the use of modern computing power, proteins can be modeled at interfacial sites with a fair degree of accuracy, and the results of such models can be directly compared to experimental SFG spectroscopy results.

The aim of this study is to further develop the practical applications of simulated SFG spectroscopy by exploring results generated for protein secondary structures. Specifically, the structures of several dozen model α-helix and β-sheet structures were simulated using a molecular dynamics program, and then the SFG response was calculated from simulated structures using a semi-empirical model established by Roeters *et al*[1]. Similar analyses on the applications of SFG spectroscopy with interfacial proteins have been performed by several different groups across the world in an attempt to better understand how proteins function[1-11]. The data generated in this study will provide a general α-helix and β-sheet response that can be used as a set to which experimentalists can compare their data. Moreover, slight differences in the responses of the structures studied herein will provide some clues as to why certain deviations from an ideal spectrum appear. The method and the information presented in this study will allow for more

meaningful conclusions on the functional mechanisms of *in situ* interfacial proteins to be gleaned from SFG spectroscopy results. A more in-depth discussion on the mentioned topics can be found in the following sections.

Protein Structures and Functions

The functionality of a protein is determined not only by its chemical composition, but by the structure and shape the protein exhibits as well. Proteins twist and fold due to intermolecular and intramolecular forces within the protein and with the solvent environment. Even a slight change in how a protein is folded can completely negate its functionality causing it to be known as denatured. Proteins consist of amino acids that are linked together via the bonding of amino groups ($–NH_2$) with carboxylic-acid groups (–COOH) called a peptide bond. As such, proteins are commonly referred to as polypeptides. There are 20 different amino acids that all contain the above-mentioned backbone structure coupled to side chain R-groups. Each sidechain has different chemical properties, such as being hydrophilic, hydrophobic, polar, nonpolar, or charged. The sequence of amino acids and the subsequent forces generated by the backbone and sidechains of these amino acids dictates the exact shape and structure of the protein. Proteins have four classified levels of structure: primary structure, secondary structure, tertiary structure, and quaternary structure[12-13].

The primary structure of a protein is simply the sequence of amino acids that make up the protein. These amino acids, as mentioned above, differ in chemical and physical properties depending on their side chain –R groups. Notably, the side chains vary in polarity which gives rise to various intermolecular and intramolecular forces. A protein's amino acid sequence typically ends with a –COOH group and an $–NH_2$ group, called the

C-terminus and N-terminus respectively[12-13]. The primary structure of a protein is a decent starting point to gaining understanding about the molecule, but the functionality of the protein is provided by the shape that the sequence induces. A protein's secondary structure is defined by the local conformations within the protein backbone. The two most common secondary structures are known as the α-helix and the β-sheet[12-13], and as such are the primary contributors to the response from proteins when using SFG spectroscopy.

The α-helix is formed by the hydrogen bonding between C=O and N-H contained throughout the amino acid backbone. These forces cause the protein backbone to take on a spiral configuration where each turn in the helix involves four amino acid residues where the oxygen in the C=O group of the first residue bonds with the hydrogen in the N-H group four residues sequentially after (i + 4 → i hydrogen bonding), though it is more accurate to say each turn is 3.6 residues long[6, 12-14]. This configuration repeats a number of times to form an α-helix. Another important characteristic about the α-helix is that the amino acid side chains are on the outside of the structure. It should be noted that other helixes can form, such as the 3_{10}-helix (i + 3 → i) and π-helix (i + 5 → i), but these structures are not as abundant as their α-helix counterpart[6, 12-14].

β-sheets get their structure from the formation of hydrogen bonds that between C=O and N-H groups in different sections of the protein backbone. This structure can form between strands running in the same direction (parallel) or with strands running in the opposite direction (anti-parallel). The anti-parallel β-sheet formation is more stable than its

parallel counterpart due to the hydrogen bonds lining up more efficiently in that formation, however neither structure is considered unstable[7, 12-13]. Like α-helixes, β-sheets form with the –R group sidechains facing outside of the backbone structure.

The tertiary structure of a protein refers to how the entire shape of the molecule conforms in 3-dimensions. This level of structure is heavily dependent on the interactions of the amino acid side chains and is not as rigid as the previous two structural levels. This structure is partially dependent on the medium that the protein is in. A protein in a polar solvent will have a different tertiary structure than the same protein in a non-polar solvent. Additionally, a protein molecule in a vacuum would exhibit far less folding due to the absence of the interactions between the protein molecule and the solvent molecules. Most proteins exist in a watery environment where hydrophobic –R groups tend to fold towards the center of the overall protein structure where they are partially shielded from the polar water molecules by the more hydrophilic –R groups. Furthermore, bonds can form between certain side chains forming a bridge. A bridge connects two parts of the protein with the formation ionic bonds (as in disulfide bridges), hydrogen bonds, or through ionic dipole charges (as in salt bridges). Interactions between the protein and chemical entities other than the solvent can affect a protein's tertiary structure. These interactions can have a significant impact on a protein's ability to carry out its immediate function such as transport proteins that absorb and deposit minerals across a membrane. The introduction of certain ions and/or compounds causes the protein to change shape to either allow or close passage

through a particular membrane[11]. An array of conformations can be observed and are outlined by one of two common conventions referred to as CATH and SCOP[12-13].

Finally, the quaternary structure of a protein refers to the overall shape of the molecule through the arrangement of the different sections, known as chains, that may be present in the protein. These chains can be identical (e.g. a homodimer of two proteins) or different (e.g. a heterodimer of two proteins) and are distinguished by groupings distinguished by the folding in the tertiary structure. The driving forces behind oligomerization are often the same as seen with the tertiary structures (salt bridging, hydrogen bonding, hydrophobic interactions)[12-13]. This quaternary structure is the final level in the intricate and complex structures of proteins.

Protein Characterization and Vibrational Spectroscopy

Common methods for characterizing protein structures include nuclear magnetic resonance (NMR), cryo-EM, and X-ray diffraction, however these methods require that proteins to be in unnatural environments and in unnatural conformations (i.e. crystalline, highly concentrated). Because of this, these methods are mostly limited to characterizing the primary structures of proteins. Protein secondary structures can be characterized relatively *in situ* using vibrational spectroscopy techniques such as Fourier transform infrared (FTIR), Raman, and SFG spectroscopies. These methods are inherently non-invasive and non-destructive which makes them ideal for analyzing sensitive compounds like proteins[1-2, 4-5, 8-11, 15]. These techniques involve irradiating a compound with a high-powered light source, typically a laser, and measuring the response. FTIR utilizes infrared light; Raman uses visible light; SFG uses a combination of IR and visible light (see figure 1). As the light strikes a protein, a photon is absorbed causing an excitation. When this excitation relaxes, a photon is released with a slightly different frequency than the absorbed photon. This difference in frequency is captured by a sensor, and provides information on what is present in the protein's structure.

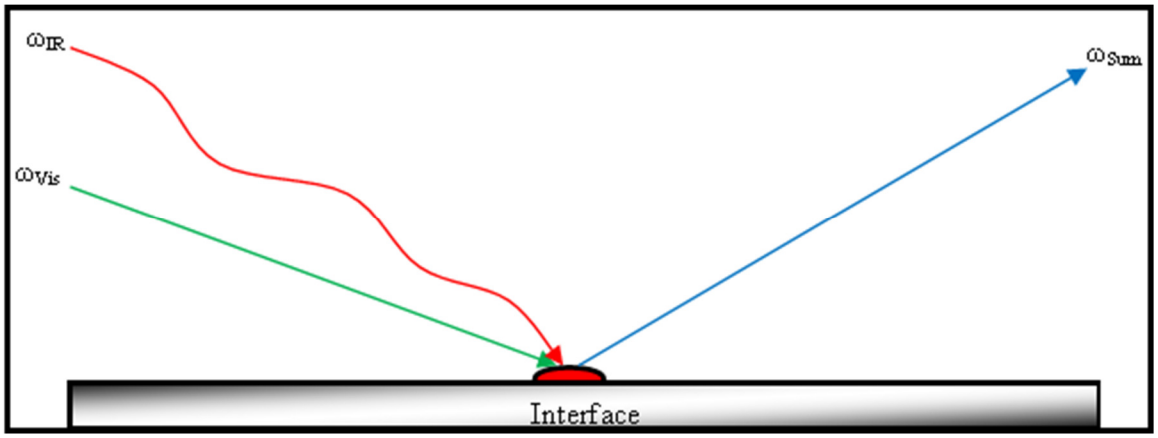

Figure 1: Representation of SFG spectroscopy with SSP polarization where ω_{IR} is the frequency of the P polarized IR beam, ω_{Vis} is the frequency of the S polarized visible light beam, and ω_{Sum} is the SFG response equal to $\omega_{IR} + \omega_{Vis}$ with S polarization.

Responses are generally recorded within certain modes of the full spectrum. The most common mode for analyzing proteins, and the mode of interest for this paper, is the amide I mode. This mode is centered around 1650-1660 cm^{-1}, and is produced mostly from the stretching of C=O bonds which comprises a large portion of protein backbones. The response at this frequency can be affected by hydrogen bonds and other dipole interactions such as those found in the secondary structures of proteins[1, 4, 6-7, 10-11, 15]. The magnitude of splitting in spectroscopy responses is determined by the distance and orientation of hydrogen-bonds and transition-dipole interactions[1, 4, 6-7, 10-11, 15]. This allows for the orientation and relative position of certain groups to be gleaned from the responses generated by spectroscopy techniques, namely SFG spectroscopy. This gives rise to being able to determine the secondary structure[1, 6-7, 10]. Responses for α-helix structures tend to form a high intensity peak near 1650 cm^{-1} that is affected by the hydrogen bonding within

the structures[1, 6, 10-11]. Responses for anti-parallel β-sheet (the more commonly formed β-sheet[12]) structures will typically consist of two peaks, one around 1620 cm^{-1} and one at a higher frequency around 1690 cm^{-1} wavenumber[10-11]. A comparison of peak intensities is used to determine the relative population of and the predominant orientation of different secondary structures[1, 4, 6-7, 9-11, 15].

There are however, limitations to the use of IR and Raman spectroscopy, mainly being that neither can be used for *in situ* analysis of interfacial proteins. This is due to two reasons: the first being that high concentrations of a given protein which is not always a natural environment for proteins and secondly, responses are generated by molecules in the bulk solution which shields any response that could be given by the minority interfacial-proteins. SFG spectroscopy does not have these limitations[1, 4, 6, 8-9, 11]. SFG spectroscopy utilizes polarized IR and visible light in specific combinations of polarization labeled as senkrecht (S) (German word meaning perpendicular) and parallel (P). The polarization is labeled with respect to the lab-frame (x, y, z) described in the next section. Different combinations result in different SFG responses[1, 6-7, 9-11], and thusly must be accounted for. What gives SFG spectroscopy the ability to see proteins at an interface *in situ* has to do with how the response is generated. Typically, a short laser pulse, on the scale of femtoseconds, is used for the visible light component, and a broader and longer pulse is used for the IR component[2, 10]. When the two beams of light strike the protein, photons are absorbed generating an excited state that's proportional to the energy from the absorbed photon. The energies from both light sources combine to cause an elevated state of

excitement which, when relaxed, generates a photon at a new frequency resulting from the summation of the original two frequencies of the IR and visible lights. This is shown above in figure 1.

Inherently, an SFG response *cannot* be obtained in media with inversion symmetry such as solvated proteins within a single medium. Therefore, it is an ideal technique to study interfacial proteins while ignoring effects from solution state proteins. Interfacial proteins play a crucial role in many biologically, medically, and industrially relevant processes. For example, proteins mediate the growth of bone and other inorganic structures in organisms ranging from the microscopic scale to humans. The proteins of interest will exist at some kind of interface: harmful proteins can adhere to medical implants to induce the formation of biofilms that can lead to infection, proteins are involved in the formation of foams at the air-water interface in fermentation processes, transport proteins facilitate the transport of minerals and nutrients across cell membranes[11]. Therefore, the ability of SFG to analyze interfacial proteins *in situ* is of particular interest to scientists. While SFG spectroscopy is a widespread method for probing protein structure and orientation at interfaces, the interpretation of experimental results is not always straightforward. The spectroscopic signal lacks molecular level detail, but with molecular simulation and careful analysis, experimental structures can be connected to reasonable, hypothetical molecular models.

Molecular Dynamics Simulations

Molecular simulations have become a powerful tool made feasible through modern computer processing power. Engineers and researchers can use simulations to gain insight into hypothetical systems/processes thus reducing the need to spend time and resources on rigorous experimentation. For the case of investigating the behavior of complex organic molecules, such as proteins, molecular dynamics (MD) simulations are used. MD simulations calculate the potential energy of a system of atoms through the integration of Newton's equations of motions through time and space. The atoms are assumed to behave like classical bodies that are acted upon by a set of inter and intra molecular forces (covalent bonds, Van der Waals, hydrogen bonds, etc.) that are defined in what's called a force field. Starting with initial values for the positions and velocities of each atom, temperature, pressure, and the force field, the total energy of the system can be calculated. Snapshots of the system are taken periodically for later analytical and visual use. This time step needs to be small enough that holding the velocities and forces constant does not create an unrealistic assumption, but large enough as to not create too many calculations for the computer to complete in a reasonable time-frame. Typically, these time steps are on the magnitude of a couple of femtoseconds. Because of the characteristic timescale of molecular motions, an integration timestep of 2 femtoseconds is the standard choice. Using

this timestep with current computing power allows for tens of nanoseconds to be simulated on a per day basis for systems of small proteins solvated in water[16-17].

The first MD simulation was performed in 1950's where the motions of a box of 32 and 108 liquid molecules were simulated with the assumptions that each molecule behaved like a hard sphere. At the time, this simplified model took 30-40 hours on the fastest contemporary compute. Today, a similar simulation can be performed using billions of molecules[17]. However, the hard-sphere model does not give accurate predictions for many situations, and with considerable interest in more complicated molecular systems more advanced models are needed. As computing technology and simulation algorithms advanced, more complex and detailed models began to be used. By the 1960's the Lennard-Jones model for molecules began to be utilized[17], and as time progressed, more advanced techniques, such as particle mesh Ewald summation (PME) and particle-particle-mesh were introduced[16-18].

In modern molecular dynamics, development of more accurate models focus primarily on algorithms. Current algorithms are based around the Verlet method for integrating the equations of motion. This is largely due to Verlet integration being light computationally while still maintaining a high degree of accuracy[17, 19]. One of the main advantages to using this method of integration is that the error in the calculation does not accumulate with time as it does with other integration methods such as the Euler method[19]. However, the Verlet method does require a "startup" function in order to obtain initial

values for positions and velocities making the initial conditions of a simulation vital to the overall accuracy. Various forms of the Verlet method are used for different scenarios such as the Beeman method which is tailored to handle larger numbers of particles. The Runge-Kutta and leapfrog integration techniques are also used to a lesser extent[17].

Despite modern computing power and relatively light weight integrating methods, further assumptions are needed to reduce the total number of calculations needed at each time step. One such assumption is the nearest-neighbor assumption. This assumes that any given atom is acted upon only by other atoms within a short range of it, typically in the range of about 1.0 to 1.5 nm, or that atom's "nearest neighbors". This assumption is valid since most forces from further away particles are largely negated by shielding from closer particles[17]. When using this assumption though it is vital to be able to accurately track the nearest neighbors of atoms which can become immensely laborious with macroscopic changes in the system such as those that may arise from concentration, pressure, and/or temperature gradients. Additional algorithms are used to calculate the impact of forces from further molecules such as particle-particle-particle mesh and PME where the atoms are constrained by a grid, and the forces acting upon those atoms are given a weight depending on their proximity. Each cell in the grid is calculated for each time step but can be calculated in parallel with one another making the utilization of multiple processors a viable way of reducing computation times[17-18, 20]. Due to the nature of the Ewald summation, periodic boundary conditions are imposed on the simulation. This can be easily accounted for by specifying sufficiently large boundaries as to avoid interactions with

duplicates of the simulated system[20]. The most computationally extensive aspect of MD simulations however, comes from the calculation of the force fields.

There are many different force fields that have been parameterized with varying degrees of accuracy in different applications, ranges of computational complexity, along with specific strengths and weaknesses tailored for certain situations. The force fields used in the simulations for this paper were AMBER99SB-ILDN[21] and TIP3P[22]. The Assisted Model Building with Energy Refinement (AMBER) force field is commonly used in MD simulations, and is generally based on the summation of potential energies from bonds, bond-angles, bond torsion, and coulombic interactions from nonbonded atoms like other force fields[17, 23]. The main difference between the AMBER force fields and other force fields is in how it handles nonpolar hydrogens bonded to heavier atoms. AMBER force fields do not explicitly differentiate these, but instead conglomerates them into the information of the heavier atom they are bonded to[17, 23]. In previous iterations of AMBER force fields there were disagreements with the calculated torsion energy term, but that has been resolved in more recent iterations. AMBER99SB-ILDN is the third generation of the AMBER force field with improved torsional parameters in protein backbones and side chains compared to previous versions. It also includes improvements to parameters involving residues such as isoleucine, leucine, aspartate, and asparagine (ILDN) where deviations from empirical data were being observed in simulations[17, 23].

Transferable intermolecular potential with 3 points (TIP3P) is a commonly used model for water molecules that represents water as having three active sites with a point charge at each atom in addition to Lennard-Jones parameters on the oxygen atom[17, 24]. While this model is not the most accurate model for water available, it does well in simulating the bulk properties of water and its intermolecular interactions[17, 24]. TIP3P is also observed to work well with the AMBER force fields[17, 23]. More advanced models that use five (TIP5P) or six active sites on the water molecules do give more detailed and accurate representations of water, but at the cost significantly increasing the necessary calculations in a simulation. For this reason, the TIP3P model is most commonly used in MD simulations involving water as a solvent for other molecules such as proteins[17, 24].

SFG THEORY

Sum frequency generation is a second-order, nonlinear, optical spectroscopy that utilizes polarized IR and Raman signals that combine to create a new signal. The frequency of this new signal is a sum of the incoming IR and Raman signals. The utility of SFG lies in its ability to ignore signals from bulk solutions which allows for observation of interfacial structures. Theoretical models for SFG have been developed and described through the efforts of several groups[1, 6-7, 10]. Since the experiment described within this paper utilized an adaptation of the program written by Roeters[1], the methodology described in his works to calculate the theoretical SFG response in the amide-I spectrum will be outlined here. We used a modified version of Roeters' code, written by Marcus Schwarting. A working preliminary version of this code can be downloaded at https://github.com/meschw04/vsfg-bellerphon.

Based on the orientation and conformation of the protein, the amide-I exciton Hamiltonian can be calculated for the system. This matrix describes the vibrational energetics of the system, and the delocalized vibrational eigenmodes can be determined by solving the time-independent Schrödinger equation. These eigenmodes are then used to determine the IR and Raman responses for the protein which are then used to determine the SFG response or molecular hyperpolarizability[1, 6-7]. It should be noted that when SFG is actually performed, as in not simulated, that the intensity of the SFG response develops

over time[10], but within the simulation, the intensities of the SFG responses are normalized for easy comparison[1].

Amide-I Hamiltonian

The exciton amide-I Hamiltonian used in the SFG simulation program is based on the protein conformation stored in a PDB file. The PDB format is the primary format used to store the structural and conformational information of proteins in the Protein Data Bank. Using the atomic coordinates found within a PDB file the amide-I Hamiltonian can be created by analyzing the local modes and couplings between the atoms. Side-chain/backbone-interactions are ignored in this treatment because the amide I response is primarily affected by the secondary structure; side chain interactions have only a slight impact on SFG responses[1]. As such, the Hamiltonian takes the form seen in equation 1.

$$\begin{pmatrix} \hbar\omega_1^0 & \kappa_{12} & \kappa_{13} & \kappa_{14} & \cdots \\ \kappa_{21} & \hbar\omega_2^0 & \kappa_{23} & \kappa_{24} & \\ \kappa_{31} & \kappa_{32} & \hbar\omega_3^0 & \kappa_{34} & \\ \kappa_{41} & \kappa_{42} & \kappa_{43} & \hbar\omega_4^0 & \\ \vdots & & & & \ddots \end{pmatrix} \quad (1)$$

Within this matrix $\hbar$ is, ω_i^0 is the frequency of local mode i, and κ_{ij} is the coupling between the local modes i and j. The coupling between different modes uses two different models depending on if the two modes are nearest-neighbors ($\kappa_{i,\, j \pm 1}$ or $\kappa_{i \pm 1,\, j}$) or non-nearest-neighbors. For nearest-neighbors the coupling values are found through density

functional theory (DFT), and non-nearest-neighbor couplings are found using the transition-dipole coupling model (TDCM)[1]. Values for the nearest-neighbor DFT approach are found using a parameterized heat map of the dihedral angles (ϕ, ψ)[1, 15]. The method for finding the coupling interaction values for non-nearest-neighbors involves approximating the interactions with a Coulomb interaction between transition-dipole moments[1, 15].

$$\kappa_{ij} = \frac{1}{4\pi\varepsilon_0}\left(\frac{\vec{\mu}_i \cdot \vec{\mu}_j}{\left|\vec{r}_{ij}\right|^3} - 3\frac{(\vec{r}_{ij} \cdot \vec{\mu}_i)(\vec{r}_{ij} \cdot \vec{\mu}_j)}{\left|\vec{r}_{ij}\right|^3}\right) \tag{2}$$

ε_0 is the dielectric constant, μ_i and μ_j are the transition-dipoles of peptide bonds *i* and *j*, and r_{ij} is the vector connecting the two dipoles. Values for the displacements, charges, and charge flows were obtained from *Hamm et. al*[15]. The coordinates of the protein backbone atoms C, O, N, and H found in PDB files are used to calculate the transition-dipoles[1].

Due to the effects of hydrogen bonding, the amide I frequencies of local-modes need to be corrected through what is known as a red-shift which is a shift of 5 cm^{-1} applied to the eigenvalues[1, 6-7, 11, 15, 25]. Peptide bonds that are followed by proline residues in the amino acid chain require a shift of 19 cm^{-1} due to the nitrogen atom in the peptide bond being bound to a carbon atom in the proline ring-sidechain as opposed to being bonded to a hydrogen atom like the other residues. One option for obtaining these red-shift values is through the use of MD simulations, however an empirical formula adapted from *Hamm et.*

al[15] for the sake of the performance of the SFG simulation. The adjusted frequency for residue *i* bonded to a proline residue and all other amides that are hydrogen-bonded are shown in equations 3 and 4 respectively.

$$\omega_i^0 = \Omega^0 - \delta\omega_{proline} - \delta\omega_{HB,i} \tag{3}$$

$$\omega_i^0 = \Omega^0 - \delta\omega_{HB,\,i} \tag{4}$$

Ω^0 is the isolated local amide I mode, and $\delta\omega_{proline}$ and $\delta\omega_{HB,\,i}$ are the red-shifts used to correct the effects from proline residues and hydrogen bonds respectively.

SFG Response

Diagonalizing the excitonic amide-I Hamiltonian gives the delocalized vibrational eigenmodes which in turn gives the eigenvalues μ^{υ} of the eigenvectors $c^{\sigma\upsilon}$ of the eigenmode $|\upsilon\rangle$[1] described in equation 5 where $|\sigma\rangle$ is the localized amide-I state of the peptide unit σ.

$$\left|\upsilon\right\rangle = \sum_{\sigma} c^{\sigma\upsilon} \left|\sigma\right\rangle \tag{5}$$

The IR transition-dipoles and Raman tensors are determined from the coordinates of the peptide backbone atoms which are found in the PDB file for the protein. Roeters defines the (a, b, c)-frame used for the Raman tensors α_{ij}^{σ} for the local mode σ as the a-axis being perpendicular to the amide plane, and the b-axis being perpendicular to the a-axis and the c-axis[1].

$$\alpha_{ij}^{\sigma} = \begin{pmatrix} 0.05 & & \\ & 0.20 & \\ & & 1.00 \end{pmatrix} \tag{6}$$

The Raman tensors are transformed into the (x, y, z)-frame using the direction cosines l_x, l_y, l_z, m_x, m_y, m_z, n_x, n_y, and n_z[1].

$$\begin{aligned}
\alpha_{xx}^{\sigma} &= l_x^2\alpha_{aa}^{\sigma} + l_y^2\alpha_{bb}^{\sigma} + l_z^2\alpha_{cc}^{\sigma} \\
\alpha_{yy}^{\sigma} &= m_x^2\alpha_{aa}^{\sigma} + m_y^2\alpha_{bb}^{\sigma} + m_z^2\alpha_{cc}^{\sigma} \\
\alpha_{zz}^{\sigma} &= n_x^2\alpha_{aa}^{\sigma} + n_y^2\alpha_{bb}^{\sigma} + n_z^2\alpha_{cc}^{\sigma} \\
\alpha_{xy}^{\sigma} &= l_x m_x\alpha_{aa}^{\sigma} + l_y m_y\alpha_{bb}^{\sigma} + l_z m_z\alpha_{cc}^{\sigma} \\
\alpha_{xz}^{\sigma} &= l_x n_x\alpha_{aa}^{\sigma} + l_y n_y\alpha_{bb}^{\sigma} + l_z n_z\alpha_{cc}^{\sigma} \\
\alpha_{yz}^{\sigma} &= m_x n_x\alpha_{aa}^{\sigma} + m_y n_y\alpha_{bb}^{\sigma} + m_z n_z\alpha_{cc}^{\sigma}
\end{aligned} \tag{7}$$

By applying equation 5 the IR and Raman responses of eigenmode can be found for $i, j, k = x, y, z$[1].

$$\mu_k^{\upsilon} = \left\langle 0 \left| \hat{\mu}_k \right| \upsilon \right\rangle = \sum_{\sigma} c^{\sigma\upsilon} \mu_k^{\sigma} \tag{8}$$

$$\alpha_{ij}^{\upsilon} = \left\langle 0 \left| \hat{\alpha}_{ij} \right| \upsilon \right\rangle = \sum_{\sigma} c^{\sigma\upsilon} \alpha_{ij}^{\sigma} \tag{9}$$

$\hat{\mu}_k$ is the electric dipole operator, $\hat{\alpha}_{ij}$ is the Raman scattering operator, and μ_k^{σ} is the IR transition-dipole moment of peptide σ. With the values from equations 8 and 9, the Intensities for the IR and Raman responses can be found[1].

$$I_{IR} \propto \sum_{\upsilon} \left| \frac{\vec{\mu}^{\upsilon}}{\omega^{\upsilon} - \omega_{IR} - i\Gamma} \right|^2 \tag{10}$$

$$I_{Raman} \propto \sum_{\upsilon} \left| \frac{\alpha^{\upsilon}}{\omega^{\upsilon} - \omega_{laser} + \omega_{Stokes} - i\Gamma} \right|^2 \tag{11}$$

I_{IR} and I_{Raman} are the intensities of the IR and Raman responses respectively, ω^{υ} is the eigenvalues from the amide-I exciton Hamiltonian, ω_{IR} is the frequency of the IR field,

ω_{laser} is the frequency of the visible light laser, ω_{Stokes} is the frequency of the Stokes field, and Γ is the line width of the Lorentzian.

The SFG hyperpolarizability, or SFG response, $\beta^{(2)}_{ijk}$ of mode υ is calculated by taking the tensor product of the IR transition-dipole moment and the Raman tensor of mode υ[1].

$$\beta^{(2)\upsilon}_{ijk} = \mu^{\upsilon}_{k} \otimes \alpha^{\upsilon}_{ij} \tag{12}$$

The summation for all modes within the protein gives the hyperpolarizability for the entire protein.

$$\tilde{\beta}^{(2)\,protein}_{ijk} = \frac{-1}{2\hbar} \sum_{\upsilon} \frac{\beta^{(2)\upsilon}_{ijk}}{\omega^{\upsilon} - \omega_{IR} - i\Gamma} \tag{13}$$

Since this form of the hyperpolarizability is frequency dependent it needs to be converted into the lab (X, Y, Z)-frame. This frame is related to the (x, y, z)-frame by three Euler angles (θ, ϕ, ψ). The Euler transformation is averaged over the entire molecular orientation distribution resulting in equation 14[1].

$$\chi^{(2)\,protein}_{IJK} = N \sum \left\langle \left(\hat{X} \cdot \hat{x}\right)\left(\hat{Y} \cdot \hat{y}\right)\left(\hat{Z} \cdot \hat{z}\right) \right\rangle \tilde{\beta}^{(2)\,protein}_{ijk} \tag{14}$$

$\chi^{(2)\,protein}_{IJK}$ is the nonlinear susceptibility within the lab frame and N is the number of molecules that contribute to the response. Depending on the polarizations of the IR and Raman lasers, the nonlinear susceptibility changes. There are two types of polarization

used for SFG which are labeled S and P. The S polarization is perpendicular to the plane of incidence for the IR and Raman beams, and the P polarization is parallel to this plane. The differences between the responses of different polarization combinations stems from differences in refractive indices. These differences are corrected by multiplying the nonlinear susceptibility factor with Fresnel factors[1, 6-7]. There are four polarization combinations that result in non-zero SFG responses from nonchiral molecules: SSP, SPS, PSS, and PPP[1, 6-7, 10]. PSP, SPP, and PPS combinations also yield non-zero SFG responses for chiral molecules[1, 6-7, 10]. The SFG calculations in this paper generated responses for the SSP and SPS polarization combinations because these two combinations are the most commonly used in experimental studies[1, 10-11]. As such, only the effective nonlinear susceptibility factors for SSP and SPS are given in equations 15 and 16.

$$\chi_{SSP}^{(2)} = L_{YY}(\omega_1)L_{YY}(\omega_2)L_{ZZ}(\omega_3)\sin(\rho_3)\chi_{YYZ}^{(2)} \tag{15}$$

$$\chi_{SPS}^{(2)} = L_{YY}(\omega_1)L_{ZZ}(\omega_2)L_{YY}(\omega_3)\sin(\rho_2)\chi_{YZY}^{(2)} \tag{16}$$

$$L_{YY}(\omega_j) = \frac{2n_1(\omega_j)\cos(\rho_j)}{n_1\cos(\rho_j) + n_2(\omega_j)\cos(\gamma_j)} \tag{17}$$

$$L_{ZZ}(\omega_j) = \frac{2n_2(\omega_j)\cos(\rho_j)}{n_1\cos(\gamma_j) + n_2(\omega_j)\cos(\rho_j)} \tag{18}$$

L_{ii} is the Fresnel factor for beam j, ω_1 is the summed frequency from ω_2 and ω_3 the Raman and IR frequencies respectively, ρ_1 is the angle that the summed frequency beam is

generated from the angles of incidence of the Raman ρ_2 and IR ρ_3, n_1 is the refractive index in media 1, n_2 is the refractive index in media 2, and γ_j is the refracted angle.

These factors are then used to calculate the intensity of the SFG response I_{SFG}. Before this can be done however, we must consider the effects from off-resonant contributions on the SFG response which can come from high-intensity responses from outside of the amide-I window. With a large enough intensity, these off-resonant responses can influence the shape of the SFG amide-I response despite being outside of the spectral window[1]. The SFG intensity is thus given by equation 19.

$$I_{SFG} \propto I_{IR} I_{Raman} \left| \chi_{IJK}^{(2)\,protein} + A_{OR} e^{i\phi_{OR}} \right|^2 \tag{19}$$

A_{OR} is the amplitude of the off-resonant contribution and ϕ_{OR} is the phase of the off-resonant contribution. Further details and derivations can be found in various literatures[1, 6-7, 11, 15].

METHODS

A total of 30 proteins and their respective PDB files were obtained from the Protein Data Bank[26] for use in the SFG simulations (see table 1). 20 proteins were chosen with the stipulation that their structure was composed of at least 80% α-helices, and the remaining 10 proteins were chosen with the stipulation that their structure was composed of at least 80% β-sheets. This was done so that around five examples of the respective secondary structure could be extracted from the PDB files in a later step. All of the proteins selected were a mix of large, small, cyclic, and asymmetrical. The proteins selected are all found in homo-sapiens as well. These specifications were chosen to give a diversity of different proteins, but no selection criteria other than the composition of secondary structures was expected have any significant impact on the results of the SFG simulations. Theoretically, any protein containing either α-helix or β-sheet structures would work sufficiently well.

The dictionary of protein secondary structure (DSSP)[12] program was used to characterize the secondary structure of each amino acid in the selected proteins. For proteins consisting of mostly α-helices, residue ID's denoted with an "H" were selected in accordance with DSSP nomenclature designating that residue as part of an α-helix. Helical structures consisting of at least 10 residues were selected. Shorter structures were not used

Table 1: Table of Proteins Used. Proteins were obtained in the form of PDB files from the Protein Data Bank[26]. Residue ID numbers represent the residues that make up the desired secondary structure excised.

	Protein	Residue ID Numbers of Chosen Structure				
β-sheets	1IJQ[27]	379-392	399-414	419-434	457-481	498-514
	1U93[28]	62-75	129-150	19-38	153-182	191-210
	1WFM[29]	14-36	46-62	91-114	--	--
	2HWZ[30]	5-24	34-54	62-75	84-106	128-149
	2IPK[31]	3-26	30-43	89-112	118-139	160-179
	2ZHR[32]	6-32	61-86	94-120	184-208	341-357
	3BKY[33]	9-23	33-52	68-84	144-163	204-219
	3CDG[34]	21-37	109-126	188-208	228-249	258-272
	3DVG[35]	9-23	34-49	62-75	131-152	193-212
	3DVN[35]	9-22	33-49	62-75	147-165	193-212
α-helices	2BSK[36]	16-32	43-71	188-272	119-148	161-176
	3BUA[37]	--	--	--	--	--
	3CEQ[38]	217-232	238-252	256-273	382-379	446-461
	3L8I[39]	34-54	70-83	98-115	124-148	158-184
	3WWV[40]	159-205	--	--	--	--
	4B18[41]	96-111	141-154	183-197	226-240	437-460
	4CQO[42]	1848-1865	1870-1885	1892-1915	1926-1948	1957-1981
	4F9K[43]	46-68	27-41	--	--	--
	4HNM[44]	156-168	179-190	--	--	--
	4JJY[45]	415-426	440-471	486-515	550-574	--
	4P39[46]	679-702	722-740	--	--	--
	4QMJ[47]	874-891	955-972	992-1005	--	--
	4QOB[48]	17-27	49-60	62-75	--	--
	4XA1[49]	27-43	1173-1238	211-230	--	--
	4YYH[50]	24-38	82-99	72-92	101-125	144-169
	5WW9[51]	12-40	45-55	72-92	101-125	144-169
	5JTI[52]	1073-1088	--	--	--	--
	5JVR[53]	10-46	55-65	--	--	--
	5N7K[54]	452-487	500-517	520-549	--	--
	5XBY[55]	9-23	28-41	--	--	--

because of increased error in calculating SFG responses of structures consisting of fewer than 10 residues[1, 6-7]. Proteins consisting of mostly β-sheet structures were given the same treatment with the exception that residues denoted by a "G", DSSP nomenclature for residues comprising a β-sheet structure, were selected. Five secondary structures were selected for each protein when possible. Some proteins did not contain at least five of the desired type of secondary structure, or they did not contain five unique secondary structures as was the case for some of the cyclic proteins. In total 64 α-helices and 48 β-sheets were selected from the 30 proteins. Once the residue ID numbers were obtained for the secondary structures, the PDB files were opened using the VMD program, and the chosen secondary structures of at least a length of 10 amino acids were excised and written to individual PDB files. A handful of these files were viewed in VMD to verify the files were created properly and did indeed contain only the desired secondary structure. SFG responses are reported to be affected by the orientation they are found in[1, 6, 9]. As such, the structures were then rotated about the x-axis from θ equals 0 to 180 degrees in 15-degree increments (see figure 2).

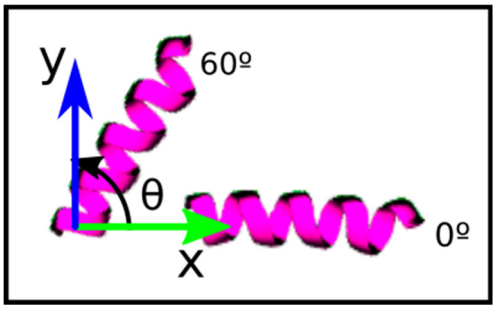

Figure 2: Rotation of α-helix Structure

The peptides containing individual secondary structures were solvated in a box of water where the edges of the box were at least 1 nm away from the solvated protein (see figures 3 and 4). This is necessary due to the periodic boundary conditions used in the simulation. If the edges were too close to the ends of the protein, then the protein would experience unnatural interactions with itself. The 1 nm distance is standard for this kind of simulation and provides a buffer zone of at least 2 nm for the protein which is sufficient to ensure the protein does not interact with itself in unnatural ways[11, 56-57]. Certain protein sidechains are charged, and thus when not balanced by oppositely charged sidechains, they can cause an overall net charge for the system. This net charge is incompatible with the PME method used for electrostatic calculations. To counter this problem, either sodium or chlorine ions are added to make the solution neutrally charged. The solvated proteins were then processed through steepest-descent energy minimization simulation in a water solvated system for 1000 steps using an open source software developed by Hamm[15]. Finally, the proteins were equilibrated using NPT (isobaric-isothermal) conditions with a reference temperature of 300K and a pressure of 1 bar for 100 ps. The AMBER99SB-ILDN and TIP3P force fields were used for this simulation as they have been shown to give relatively accurate results for this kind of system[11]. The equilibration simulations allowed for the rearrangement of hydrogen bonding both internally and with respect to the solvent. Because the initial structures were determined from crystal structures, the hydrogen bond arrangement can change dramatically. The equilibration step would ensure that there are less errant structural dynamics in subsequent simulations, and that the proteins are in a

more natural state. Both of these will help reduce error in the SFG calculations with respect to experimental observations in case we work with experimental collaborators in the future.

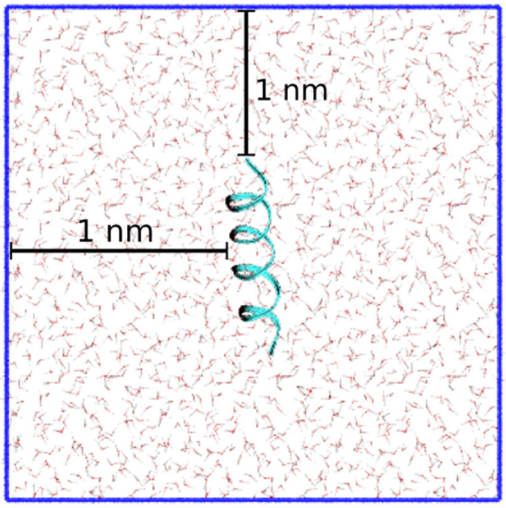

Figure 3: Representation of α-helix Solvated in Water. Boundaries of the system box are at least 1 nm from the peptide on all sides to prevent unwanted interactions.

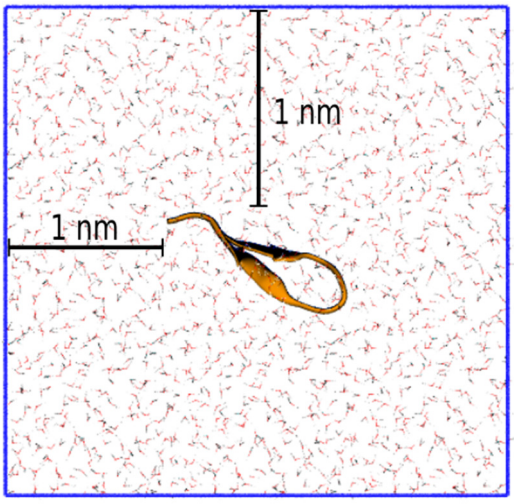

Figure 4: Representation of β-sheet Solvated in Water. Boundaries of the system box are at least 1 nm from the peptide on all sides to prevent unwanted interactions.

To calculate the SFG response of the peptides, excitonic Hamiltonian matrices were then generated for all of the prepared secondary structures using GROMACS[56-62]. These were then fed into a version of Roeter's software[1] that was adapted for use in Python by Marcus Schwarting. This program calculated the expected SFG responses for each of the secondary structures which were then analyzed and compared using a juPYter notebook. The formulation for calculating the SFG responses is explained in the *SFG Theory* section.

RESULTS AND DISCUSSION

General

The SFG responses were calculated successfully except for peptides containing proline residue(s). At the time of writing this paper, the program used for the SFG simulations was not able to resolve the unusual bonding between sidechains and backbone that occurs in proline residues. Future iterations of the program will include this functionality. As such, the SFG responses for these peptides were omitted from the analysis even though their structures were probed using MD simulations. These simulated structures are available for future analysis when proline functionality is available. It should also be noted that the plotted intensity on all of the graphs presented within have been normalized to a value of one with arbitrary units, unless otherwise specified.

Figure 5 shows the results for an α-helix and β-sheet structure across 180 degrees of rotation with the SSP polarization configuration. Looking at the plots of the various SFG responses makes it immediately obvious that not only the location of the primary peak, but the shape of the response changed with the orientation of the protein structures. This aligns with previous results reported in literature[1, 6-8, 10-11]. Some SFG responses appeared as a singular peak whereas others manifested as multiple peaks. These results are represented to some extent across all calculated SFG responses. Another interesting detail shown in these plots is that there is a symmetry of responses around 90 degrees of rotation. There is however some variation between responses that should be theoretically the same. This is likely due to some measure of error that is inherent in MD simulations where atomic

positions are not stored with precision higher than 0.001 nm. Additionally, rotations were performed in GROMACS, with respect to the x-axis normal to the surface, on these structures with low resolution, thus adding more truncation error[16-17]. Our experimental collaborators in the Weidner group (Aarhus University, Aarhus, Denmark) have reported skepticism from other spectroscopic groups, whom do not use SFG spectroscopy, on the ability of SFG spectroscopy to differentiate between unique structures at varying orientations. The results of the calculations presented within demonstrate the ability of SFG spectroscopy to differentiate between proteins at various states of conformation and orientation in combination with MD simulations.

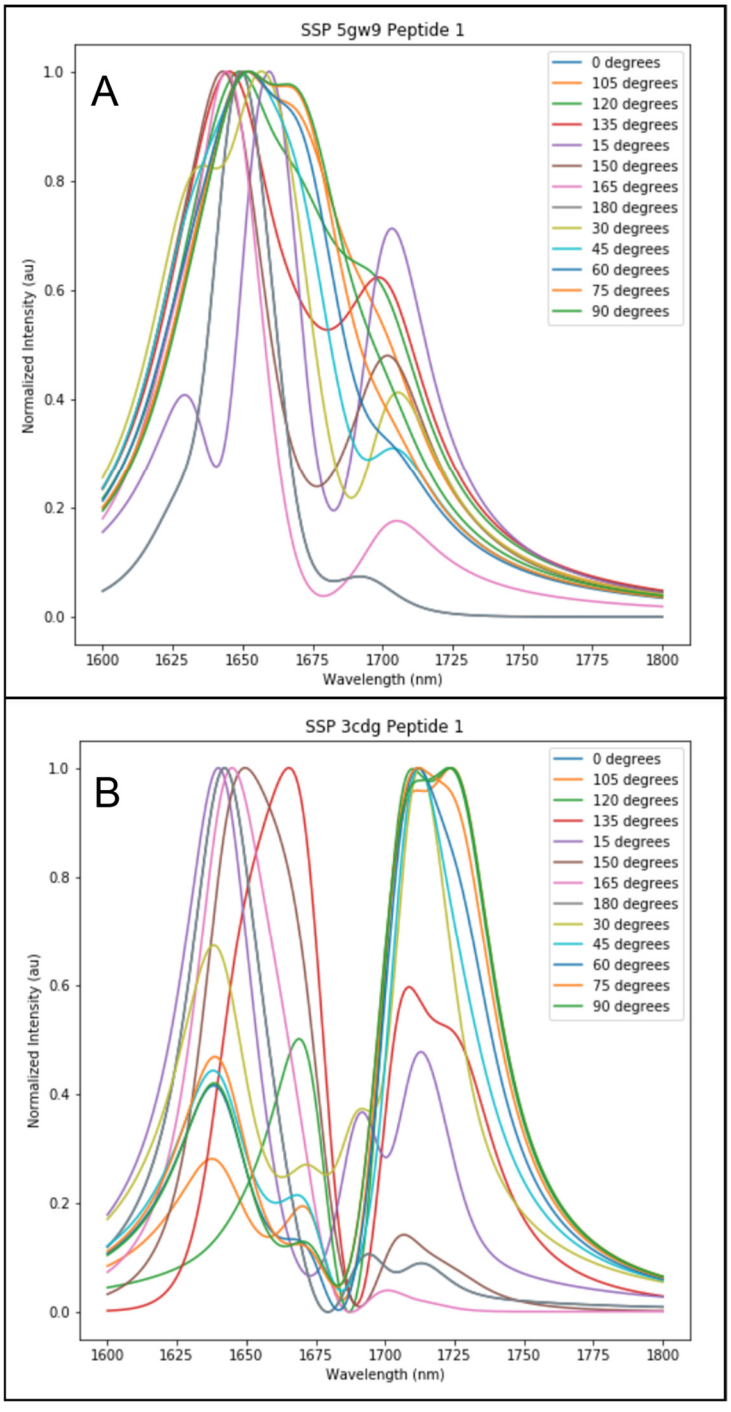

Figure 5:SSP SFG Spectra Across 180 Degrees of Rotation for A) an α-helix and B) a β-sheet.

Formation of "Shoulders" in SFG Response Averages

Responses were grouped by structure, polarization combination, and orientation, and an averaged peak was then determined for each group along with a standard deviation. Figures 6 and 7 show the SFG responses for α-helical structures using an SPS polarization at 0 degrees and 60 degrees respectively. Looking at the two averaged peaks shows the difference in the typical formation of responses at the two orientations. The alpha-helices oriented at 0 degrees gave an average peak response at 1668 cm^{-1} with a slight shoulder forming around 1685 cm^{-1} whereas those oriented at 60 degrees gave an average peak response at 1656 cm^{-1} with a more pronounced shoulder forming around 1700 cm^{-1}. With the way the average peaks were calculated, secondary peaks are not properly shown in the average. Instead of showing up as a distinct peak followed and preceded by a valley, these secondary peaks show as a shoulder or protrusion off the main peak. However, the presence of a shoulder on the averaged peak does not necessarily indicate a large presence of a lower intensity secondary peak. Shoulders can also form from a cluster of main peaks located further away from average peak in the cluster that makes up the main peak. To determine which shoulders are formed by secondary peaks or by clusters of peaks, one needs to look at the individual SFG responses that form the average. This is not immediately clear however when looking at figures 6 and 7. The shoulder formed in figure 6 appears to be due to a cluster of peaks away from the main grouping, and that the shoulder seen in figure 7 is mostly the result of secondary peaks. Looking at plots for beta-sheets structures gives a little bit clearer image of the main contributors to the shoulders formed by the averages

due to the fewer number of results and subsequent plots present. Figure 8 shows a similar plot as those mentioned above, but for a beta-sheet oriented at 30 degrees generated with an SSP polarization combination. For this average, the shoulder appears to be mainly generated from secondary peaks although there is some contribution from primary peaks evident. In general, it appears that shoulders generated by clusters of primary peaks are closer in intensity to the primary average peak than shoulders generated by clusters of secondary peaks which appear at lower intensities.

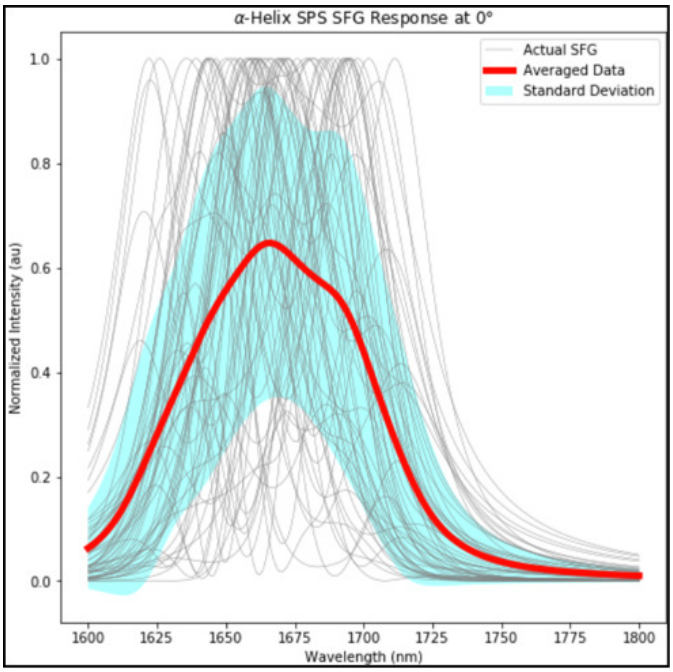

Figure 6: Averaged SPS SFG Spectra for α-helices Oriented at 0 Degrees.

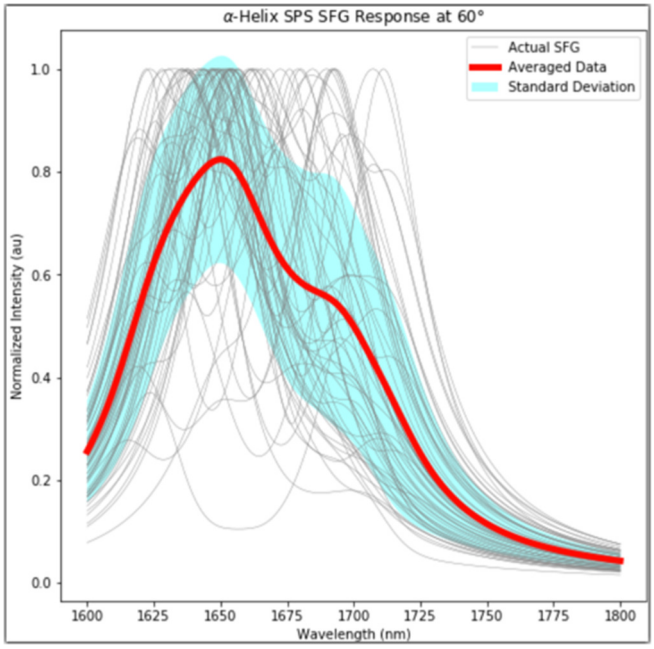

Figure 7: Averaged SPS SFG Spectra for α-helices Oriented at 60 Degrees.

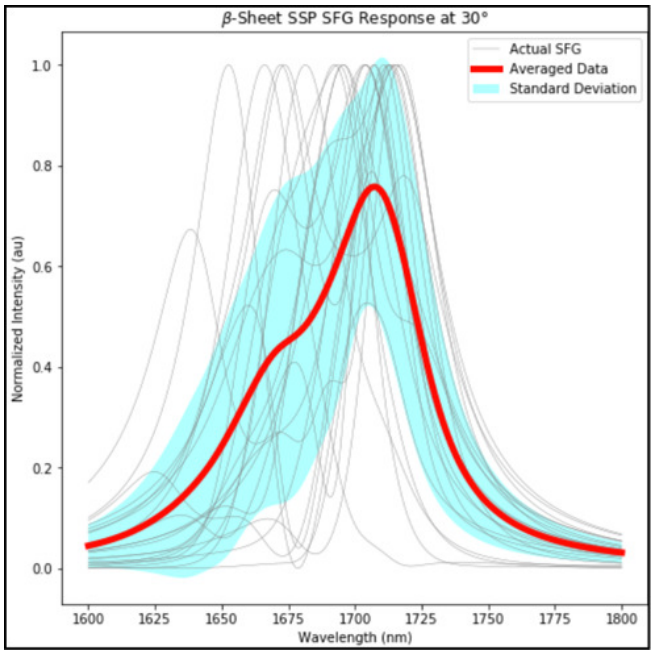

Figure 8: Averaged SSP SFG Spectra for β -sheets Oriented at 30 Degrees

Effects of Orientation and Polarization on Averaged Responses

As stated previously in this paper, and by many researchers studying applications of SFG spectroscopy, SFG spectroscopy provides a more detailed response than either Raman or IR spectroscopy individually[1, 6-8, 10-11]. One such detail that SFG can describe is the orientation of a molecule which was shown in the figures presented prior in this section. The averaged responses described in the previous section, also show some of the variations in SFG responses caused by changing the orientation of the molecule.

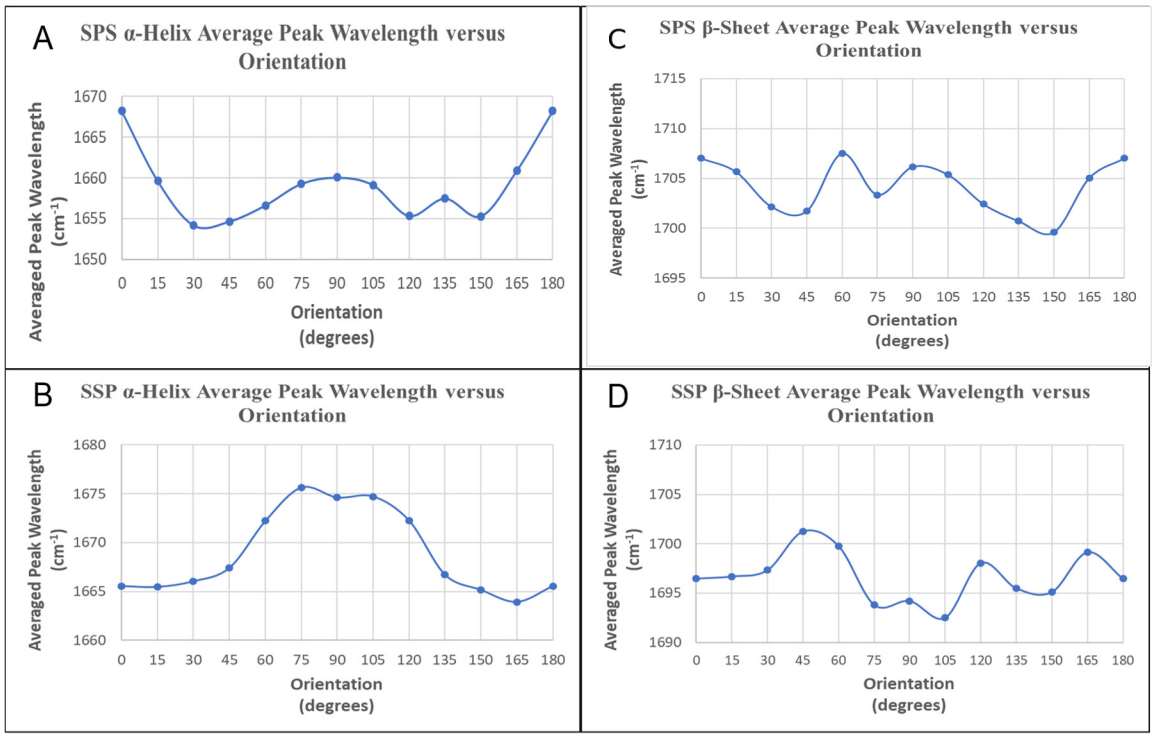

Figure 9: Average Peak Wavenumber with Respect to Orientation A) Averaged responses of α-helix structures from SPS SFG over 180 degrees of rotation, B) Averaged responses of α-helix structures from SPS SFG over 180 degrees of rotation, C) Averaged responses of β-sheet structures from SSP SFG over 180 degrees of rotation, D) Averaged responses of β-sheet structures from SSP SFG over 180 degrees of rotation

Figure 9 shows plots of the averaged peaks for the results of the two secondary structures at SPS and SSP SFG polarization. The symmetry around the 90 degrees orientation is shown well with the α-helix responses, but that same symmetry is not represented by the averaged responses for the β-sheet structures. Potential causes for this are discussed in the next section. For now, observations will be made using mainly the responses from the α-helix structures.

Apart from the expected symmetry, figure 9 shows that there does appear to be a shift in the location of the primary peaks as the orientation of the structures is changed. It should be noted however, that the entirety of the change from the two furthest averages still fall well within the about 20 cm^{-1} standard deviation found for all the averaged responses. This standard deviation likely comes from slight curves/deviations from "ideal" secondary structures and from error introduced by the MD simulations which was described earlier[1, 4-11, 16-17, 25]. The variations seen all fit within this 20 cm^{-1} standard deviation, and as such could be attributed to the error represented described by that standard deviation. The location of the main peak is also not the only difference in the SFG responses that manifested. The addition/subtraction of one or more secondary peaks, and even multiple primary peaks, was observed within the responses of the same structure at different orientations. A similar effect was also seen between differing polarization responses. Figure 10 shows a comparison of averaged responses for the two secondary structures at 0, 30, 60, and 90-degree orientations using SPS and SSP polarizations.

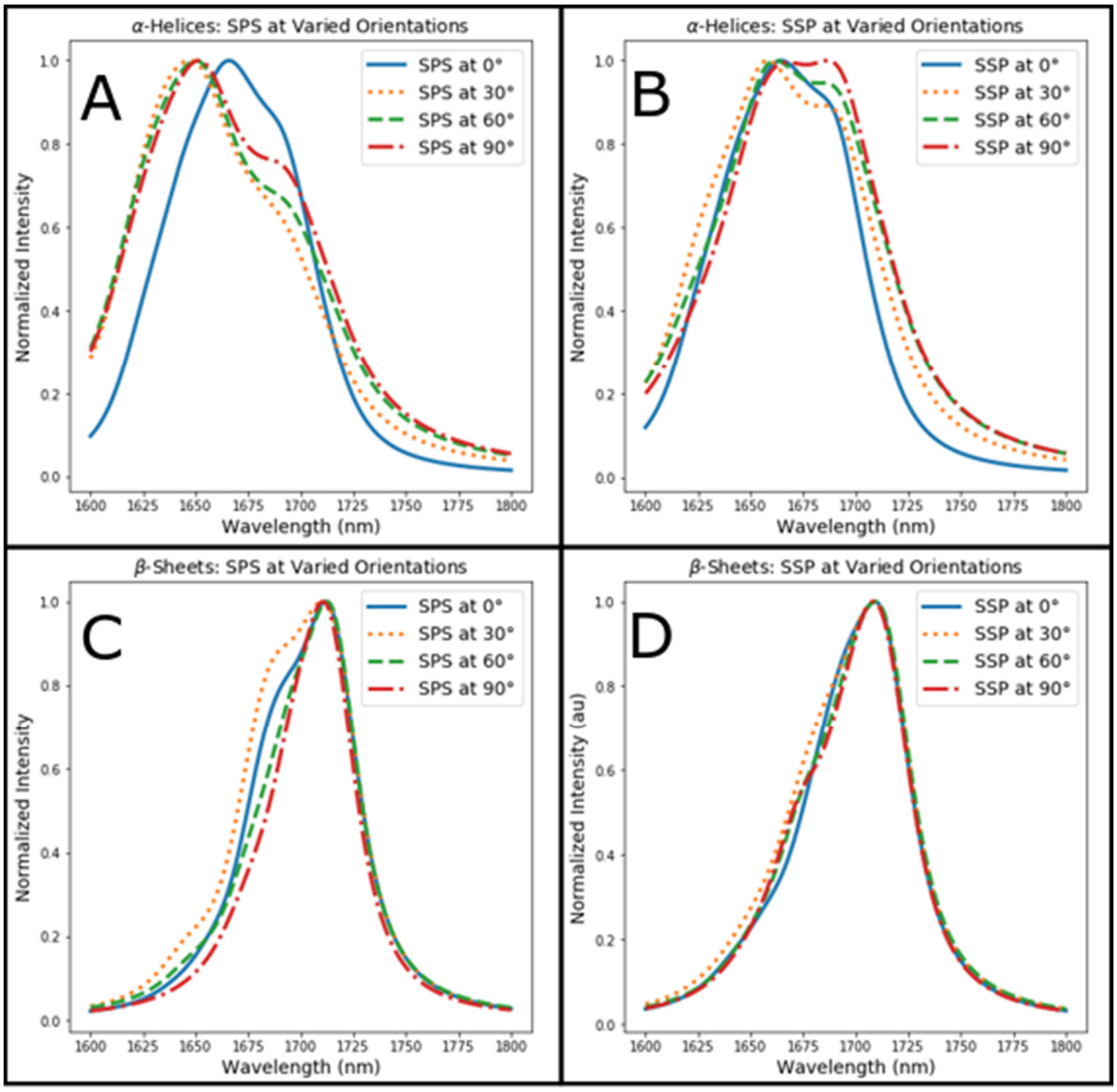

Figure 10: Comparison of Averaged Responses for Different Orientations A) Shows the averaged responses for α-helix structures using SPS polarization, B) Shows the averaged responses for α-helix structures using SSP polarization, C) Shows the averaged responses for β-sheet structures using SPS polarization, D) Shows the averaged responses for β-sheet structures using SSP polarization

In this representation, the shift of the average peak wavenumber due to changes in orientation appears miniscule, however it is easier to see the changes in shape with regards to orientation. As mentioned before, the averaged responses are not ideal for showing multi-peaked responses as those tend to manifest as "shoulders" in the averaged responses as opposed to actual peaks. However, this is still enough to clearly see that there is variation in the shape of the SFG response when the orientation of the molecule is changed. Changes to the shape of the responses can also be seen when the polarization used in for the SFG analysis is changed. This phenomenon is expected because of the contribution that the polarization has in the theoretical SFG calculations seen in equations 15 and 16. It has also been reported several times before[1, 6-7, 10-11, 25]. However, there is again little variation seen amongst the responses for the β-sheet structures.

Investigation into SFG Responses from β-sheet Structures

Further investigation into these structures revealed that the hydrogen bonds that held the protein backbone in the β-sheet conformation were no longer present after a short MD simulation. The PDB file for these structures showed that the peptides were still in the general shape of a β-sheet, but the segments of protein backbone were further apart than the structures that maintained their original shapes (see figure 11). Considering a large number of the β-sheet responses were not able to calculate due to containing proline residues, and that there were fewer of these structures selected to begin with, the averaged

responses were considerably skewed by these β-strand structures. The main identifier of the β-strand structures was the shape of the SFG responses which showed as a singular (or mostly singular) peak (see figure 12). This is likely because of the effects that hydrogen bonds have on SFG responses[1, 6-7, 10-11, 25]. Since the hydrogen bonds in the β-sheet structures are not present in the β-strand structures, the SFG responses show almost exclusively the response from the amide bond in the backbone of the peptides absent from the redshift caused by hydrogen bonding networks. The location of the peaks further supports this explanation. SFG responses for proteins consisting of mostly β-sheet structures show two large peaks around 1620 and 1675 cm^{-1} wavenumbers[1, 4-5, 10-11, 25]. The SFG responses presented within this paper tend to be closer to 1700 cm^{-1} and have only one peak. This indicates that the shape and location of SFG peaks is largely determined by the hydrogen bonding that takes place in different molecules.

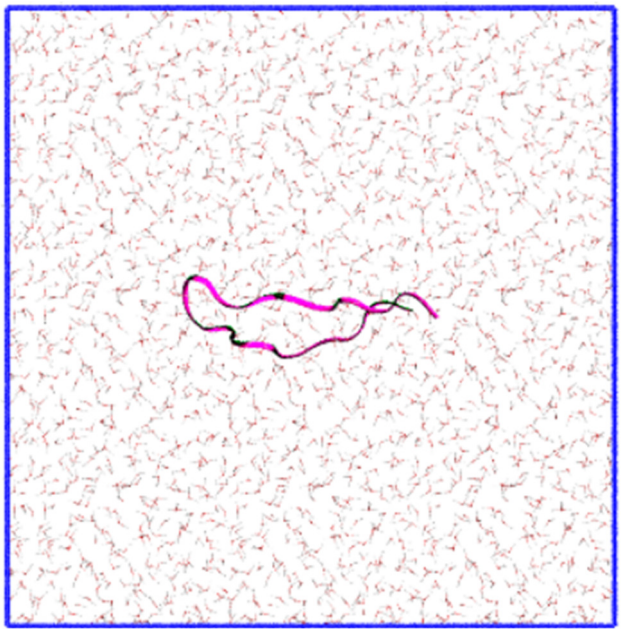

Figure 11: β-strand Structure Solvated in a Box of Water.

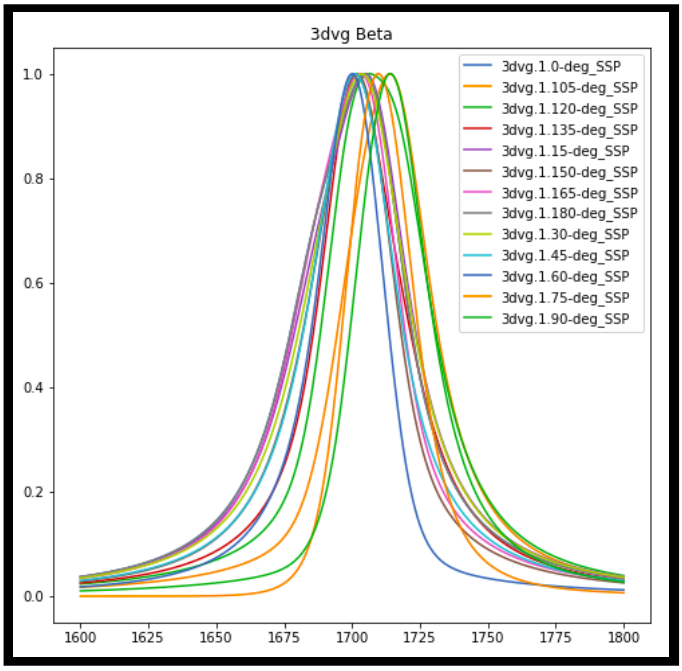

Figure 12: SSP SFG Spectra of a β-strand Structure.

CONCLUSIONS

The calculated SFG responses for the individual α-helix and β-sheet secondary structures varied based upon several factors. Averaged values for the responses were determined for the two types of secondary structures at different orientations and polarization combinations. These averaged responses represented the presence of multiple peaks within the original responses through the formation of a "shoulder" off the main peak. While the location of the primary peaks did vary with changing peptide orientation, these variations were well within the calculated standard deviation of around 20 cm^{-1}. Variations are expected to primarily come from the structures having slight curvatures/deviations from ideal secondary structures, from errors introduced by the low resolution of atomic positions in MD simulations, and from the assumptions used in the formulation of the SFG calculations. More pronounced changes were observed in the shape of the averaged responses for the α-helix structures. The averaged peaks for these structures were found at around a wavenumber of 1660 cm^{-1} which conforms with values published for SFG responses from α-helix structures. The results from the β-sheet structures however produced unusual spectra when compared to previously reported results for IR and SFG spectroscopy. The cause of this deviation is the degradation of hydrogen bonding networks during the MD simulation likely due to the change in local environment when the peptide was excised from its crystal structure. These structures, herein referred to as β-strands, gave responses primarily consisting of a singular peak at around 1690 cm^{-1} which aligns with

previously reported results for protein backbones that are without any hydrogen bonding network. The β-sheet structures that maintained their conformation gave results similar to those previously reported with two peaks at around 1630 cm^{-1} and above 1680 cm^{-1} wavenumbers. Changes to the shape of the responses were also seen respective to the use of the particular SFG polarization combination, SSP or SPS. Overall, the results demonstrate the ability to differentiate among protein secondary structures and orientations using SFG spectroscopy in combination with theory and molecular simulations.

REFERENCES

1. Roeters, S. J.; van Dijk, C. N.; Torres-Knoop, A.; Backus, E. H. G.; Campen, R. K.; Bonn, M.; Woutersen, S., Determining In Situ Protein Conformation and Orientation from the Amide-I Sum-Frequency Generation Spectrum: Theory and Experiment. *The Journal of Physical Chemistry A* **2013,** *117* (29), 6311-6322.
2. Shen, Y. R., Surface properties probed by second-harmonic and sum-frequency generation. *Nature* **1989,** *337*, 519.
3. Chen, Z.; Shen, Y. R.; Somorjai, G. A., Studies of Polymer Surfaces by Sum Frequency Generation Vibrational Spectroscopy. *Annual Review of Physical Chemistry* **2002,** *53* (1), 437-465.
4. Chen, X.; Wang, J.; Sniadecki, J. J.; Even, M. A.; Chen, Z., Probing α-Helical and β-Sheet Structures of Peptides at Solid/Liquid Interfaces with SFG. *Langmuir* **2005,** *21* (7), 2662-2664.
5. Perry, J. M.; Moad, A. J.; Begue, N. J.; Wampler, R. D.; Simpson, G. J., Electronic and Vibrational Second-Order Nonlinear Optical Properties of Protein Secondary Structural Motifs. *The Journal of Physical Chemistry B* **2005,** *109* (42), 20009-20026.
6. Nguyen, K. T.; Le Clair, S. V.; Ye, S.; Chen, Z., Orientation Determination of Protein Helical Secondary Structures Using Linear and Nonlinear Vibrational Spectroscopy. *The Journal of Physical Chemistry B* **2009,** *113* (36), 12169-12180.
7. Nguyen, K. T.; King, J. T.; Chen, Z., Orientation Determination of Interfacial β-Sheet Structures in Situ. *The Journal of Physical Chemistry B* **2010,** *114* (25), 8291-8300.
8. Liu, Y.; Jasensky, J.; Chen, Z., Molecular Interactions of Proteins and Peptides at Interfaces Studied by Sum Frequency Generation Vibrational Spectroscopy. *Langmuir* **2012,** *28* (4), 2113-2121.
9. Liu, W.-T.; Shen, Y. R., In situ sum-frequency vibrational spectroscopy of electrochemical interfaces with surface plasmon resonance. *Proceedings of the National Academy of Sciences* **2014,** *111* (4), 1293-1297.
10. Yan, E. C. Y.; Wang, Z.; Fu, L., Proteins at Interfaces Probed by Chiral Vibrational Sum Frequency Generation Spectroscopy. *The Journal of Physical Chemistry B* **2015,** *119* (7), 2769-2785.
11. Lutz, H. Shedding Light on Peptide Controlled Silica Mineralization. University of Amsterdam, 2017.
12. Kabsch, W.; Sander, C., Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **1983,** *22* (12), 2577-2637.
13. Protein Structure. Particle Sciences: www.particlesciences.com, 2009.
14. Pauling, L.; Corey, R. B.; Branson, H. R., The structure of proteins: Two hydrogen-bonded helical configurations of the polypeptide chain. *Proceedings of the National Academy of Sciences* **1951,** *37* (4), 205-211.
15. Hamm, P.; Zanni, M., *Concepts and Methods of 2D Infrared Spectroscopy*. Cambridge University Press: Cambridge, 2011.
16. Levitt, M., Birth & Future of Multiscale Modeling of Macromolecules. Stanford Department of Structural Biology & Computer Science: Georgia Tech, 2014.
17. Vlachakis, D.; Bencurova, E.; Papangelopoulos, N.; Kossida, S., Chapter Seven - Current State-of-the-Art Molecular Dynamics Methods and Applications. In *Advances in Protein Chemistry and Structural Biology*, Donev, R., Ed. Academic Press: 2014; Vol. 94, pp 269-313.

18. Wells, B. A.; Chaffee, A. L., Ewald Summation for Molecular Simulations. *Journal of Chemical Theory and Computation* **2015,** *11* (8), 3684-3695.
19. Verlet, L., Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules. *Physical Review* **1967,** *159* (1), 98-103.
20. S. W. de Leeuw, J. W. P., E. R. smith, Simulation of electrostatic systems in periodic boundary conditions. I. Lattice sums and dielectric constants. *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences* **1980,** *373* (1752), 27.
21. Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E., Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins* **2010,** *78* (8), 1950-1958.
22. Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L., Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics* **1983,** *79* (2), 926-935.
23. D.A. Case, D. S. C., T.E. Cheatham, III, T.A. Darden, R.E. Duke, T.J. Giese, H. Gohlke, A.W. Goetz, D. Greene, N. Homeyer, S. Izadi, A. Kovalenko, T.S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, D. Mermelstein, K.M. Merz, G. Monard, H. Nguyen, I. Omelyan, A. Onufriev, F. Pan, R. Qi, D.R. Roe, A. Roitberg, C. Sagui, C.L. Simmerling, W.M. Botello-Smith, J. Swails, R.C. Walker, J. Wang, R.M. Wolf, X. Wu, L. Xiao, D.M. York and P.A. Kollman, Amber 2017 Reference Manual. AMBER 2017: University of California, San Francisco, 2017.
24. Mao, Y.; Zhang, Y., Thermal conductivity, shear viscosity and specific heat of rigid water models. *Chemical Physics Letters* **2012,** *542*, 37-41.
25. Wang, L.; Middleton, C. T.; Zanni, M. T.; Skinner, J. L., Development and Validation of Transferable Amide I Vibrational Frequency Maps for Peptides. *The Journal of Physical Chemistry. B* **2011,** *115* (13), 3713-3724.
26. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E., The Protein Data Bank. *Nucleic Acids Research* **2000,** *28* (1), 235-242.
27. Jeon, H.; Meng, W.; Takagi, J.; Eck, M. J.; Springer, T. A.; Blacklow, S. C., Implications for familial hypercholesterolemia from the structure of the LDL receptor YWTD-EGF domain pair. *Nature structural biology* **2001,** *8* (6), 499-504.
28. Bryson, S.; Julien, J. P.; Hynes, R. C.; Pai, E. F., Crystallographic definition of the epitope promiscuity of the broadly neutralizing anti-human immunodeficiency virus type 1 antibody 2F5: vaccine design implications. *J Virol* **2009,** *83* (22), 11862-75.
29. *The first C2 domain of human synaptotagmin XIII*. http://www.rcsb.org/structure/1WFM.
30. *null*. http://www.rcsb.org/structure/2HWZ.
31. Venkatraman, P.; Nguyen, T. T.; Sainlos, M.; Bilsel, O.; Chitta, S.; Imperiali, B.; Stern, L. J., Fluorogenic probes for monitoring peptide binding to class II MHC proteins in living cells. *Nature chemical biology* **2007,** *3* (4), 222-8.
32. Shimizu, H.; Tosaki, A.; Kaneko, K.; Hisano, T.; Sakurai, T.; Nukina, N., Crystal structure of an active form of BACE1, an enzyme responsible for amyloid beta protein production. *Molecular and cellular biology* **2008,** *28* (11), 3663-71.
33. Du, J.; Wang, H.; Zhong, C.; Peng, B.; Zhang, M.; Li, B.; Hou, S.; Guo, Y.; Ding, J., Crystal structure of chimeric antibody C2H7 Fab in complex with a CD20 peptide. *Molecular immunology* **2008,** *45* (10), 2861-8.

34. Petrie, E. J.; Clements, C. S.; Lin, J.; Sullivan, L. C.; Johnson, D.; Huyton, T.; Heroux, A.; Hoare, H. L.; Beddoe, T.; Reid, H. H.; Wilce, M. C.; Brooks, A. G.; Rossjohn, J., CD94-NKG2A recognition of human leukocyte antigen (HLA)-E bound to an HLA class I leader sequence. *The Journal of experimental medicine* **2008,** *205* (3), 725-35.
35. Newton, K.; Matsumoto, M. L.; Wertz, I. E.; Kirkpatrick, D. S.; Lill, J. R.; Tan, J.; Dugger, D.; Gordon, N.; Sidhu, S. S.; Fellouse, F. A.; Komuves, L.; French, D. M.; Ferrando, R. E.; Lam, C.; Compaan, D.; Yu, C.; Bosanac, I.; Hymowitz, S. G.; Kelley, R. F.; Dixit, V. M., Ubiquitin chain editing revealed by polyubiquitin linkage-specific antibodies. *Cell* **2008,** *134* (4), 668-78.
36. Webb, C. T.; Gorman, M. A.; Lazarou, M.; Ryan, M. T.; Gulbis, J. M., Crystal structure of the mitochondrial chaperone TIM9.10 reveals a six-bladed alpha-propeller. *Molecular cell* **2006,** *21* (1), 123-33.
37. Chen, Y.; Yang, Y.; van Overbeek, M.; Donigian, J. R.; Baciu, P.; de Lange, T.; Lei, M., A shared docking motif in TRF1 and TRF2 used for differential recruitment of telomeric proteins. *Science* **2008,** *319* (5866), 1092-6.
38. *The TPR domain of Human Kinesin Light Chain 2 (hKLC2).* http://www.rcsb.org/structure/3CEQ.
39. Li, X.; Zhang, R.; Zhang, H.; He, Y.; Ji, W.; Min, W.; Boggon, T. J., Crystal structure of CCM3, a cerebral cavernous malformation protein critical for vascular integrity. *The Journal of biological chemistry* **2010,** *285* (31), 24099-107.
40. Yokoyama, H.; Matsui, I., Crystal structure of the stomatin operon partner protein from Pyrococcus horikoshii indicates the formation of a multimeric assembly. *FEBS open bio* **2014,** *4*, 804-12.
41. Jeong, S. A.; Kim, K.; Lee, J. H.; Cha, J. S.; Khadka, P.; Cho, H. S.; Chung, I. K., Akt-mediated phosphorylation increases the binding affinity of hTERT for importin alpha to promote nuclear translocation. *Journal of cell science* **2015,** *128* (12), 2287-301.
42. Bhandari, D.; Raisch, T.; Weichenrieder, O.; Jonas, S.; Izaurralde, E., Structural basis for the Nanos-mediated recruitment of the CCR4-NOT complex and translational repression. *Genes & development* **2014,** *28* (8), 888-901.
43. *Crystal Structure of human cAMP-dependent protein kinase type I-beta regulatory subunit (fragment 11-73), Northeast Structural Genomics Consortium (NESG) Target HR8613A.* http://www.rcsb.org/structure/4F9K.
44. Huang, X.; Wang, G.; Wu, Y.; Du, Z., The structure of full-length human CTNNBL1 reveals a distinct member of the armadillo-repeat protein family. *Acta crystallographica. Section D, Biological crystallography* **2013,** *69* (Pt 8), 1598-608.
45. Pashkova, N.; Gakhar, L.; Winistorfer, S. C.; Sunshine, A. B.; Rich, M.; Dunham, M. J.; Yu, L.; Piper, R. C., The yeast Alix homolog Bro1 functions as a ubiquitin receptor for protein sorting into multivesicular endosomes. *Developmental cell* **2013,** *25* (5), 520-33.
46. Schatz-Jakobsen, J. A.; Yatime, L.; Larsen, C.; Petersen, S. V.; Klos, A.; Andersen, G. R., Structural and functional characterization of human and murine C5a anaphylatoxins. *Acta crystallographica. Section D, Biological crystallography* **2014,** *70* (Pt 6), 1704-17.
47. Fox, J. C.; Howard, A. E.; Currie, J. D.; Rogers, S. L.; Slep, K. C., The XMAP215 family drives microtubule polymerization using a structurally diverse TOG array. *Molecular biology of the cell* **2014,** *25* (16), 2375-92.
48. *Crystal structure of cPOP1.* http://www.rcsb.org/structure/4QOB.

49. Taylor, K. C.; Buvoli, M.; Korkmaz, E. N.; Buvoli, A.; Zheng, Y.; Heinze, N. T.; Cui, Q.; Leinwand, L. A.; Rayment, I., Skip residues modulate the structural properties of the myosin rod and guide thick filament assembly. *Proc. Natl. Acad. Sci. U. S. A.* **2015,** *112* (29), E3806-15.
50. Flynn, E. M.; Huang, O. W.; Poy, F.; Oppikofer, M.; Bellon, S. F.; Tang, Y.; Cochran, A. G., A Subset of Human Bromodomains Recognizes Butyryllysine and Crotonyllysine Histone Peptide Modifications. *Structure (London, England : 1993)* **2015,** *23* (10), 1801-14.
51. Williams, S. M.; Chatterji, D., Flexible aspartates propel iron to the ferroxidation sites along pathways stabilized by a conserved arginine in Dps proteins from Mycobacterium smegmatis. *Metallomics : integrated biometal science* **2017,** *9* (6), 685-698.
52. Mariotti, L.; Templeton, C. M.; Ranes, M.; Paracuellos, P.; Cronin, N.; Beuron, F.; Morris, E.; Guettler, S., Tankyrase Requires SAM Domain-Dependent Polymerization to Support Wnt-beta-Catenin Signaling. *Molecular cell* **2016,** *63* (3), 498-513.
53. Phillips, R. K.; Peter, L. G.; Gilbert, S. P.; Rayment, I., Family-specific Kinesin Structures Reveal Neck-linker Length Based on Initiation of the Coiled-coil. *The Journal of biological chemistry* **2016,** *291* (39), 20372-86.
54. Schuetz, A.; Radusheva, V.; Krug, S. M.; Heinemann, U., Crystal structure of the tricellulin C-terminal coiled-coil domain reveals a unique mode of dimerization. *Annals of the New York Academy of Sciences* **2017,** *1405* (1), 147-159.
55. Youn, S. J.; Kwon, N. Y.; Lee, J. H.; Kim, J. H.; Choi, J.; Lee, H.; Lee, J. O., Construction of novel repeat proteins with rigid and predictable structures using a shared helix method. *Sci Rep* **2017,** *7* (1), 2595.
56. Páll S., A. M. J., Kutzner C., Hess B., Lindahl E., Tackling Exascale Software Challenges in Molecular Dynamics Simulations with GROMACS. In *EASC 2014: Solving Software Challenges for Exascale*, Springer, Cham: 2015; Vol. 8759, pp 3-27.
57. Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E., GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **2015,** *1-2*, 19-25.
58. Pronk, S.; Páll, S.; Schulz, R.; Larsson, P.; Bjelkmar, P.; Apostolov, R.; Shirts, M. R.; Smith, J. C.; Kasson, P. M.; van der Spoel, D.; Hess, B.; Lindahl, E., GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* **2013,** *29* (7), 845-854.
59. Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E., GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *Journal of Chemical Theory and Computation* **2008,** *4* (3), 435-447.
60. Spoel, D. V. D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C., GROMACS: Fast, flexible, and free. *Journal of Computational Chemistry* **2005,** *26* (16), 1701-1718.
61. Lindahl, E., Hess, B. & van der Spoel, GROMACS 3.0: a package for molecular simulation and trajectory analysis. *Molecular Modeling Annual* **2001,** *7* (8), 306-317.
62. Berendsen, H. J. C.; van der Spoel, D.; van Drunen, R., GROMACS: A message-passing parallel molecular dynamics implementation. *Computer Physics Communications* **1995,** *91* (1), 43-56.

CURRICULUM VITA

NAME: Andrew James Adams

ADDRESS: Department of Chemical Engineering
216 Eastern Pkwy
University of Louisville
Louisville, KY 40208

DOB: Dayton, Kentucky - December 16, 1994

EDUCATION
& TRAINING: B.S., Chemical Engineering
University of Louisville
2013-2017